Statistics and Probability Letters 114 (2016) 54–59

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Sharp total variation bounds for finitely exchangeable arrays

Alexander Volfovsky *, Edoardo M. Airoldi

Statistics Department, Harvard University, Cambridge, MA, USA

ARTICLE INFO

Article history:
Received 25 May 2015
Received in revised form 21 February 2016
Accepted 21 February 2016
Available online 12 March 2016

Keywords:
Finite exchangeability
Arrays
Total variation bound
Networks
Hypothesis testing

ABSTRACT

In this article we demonstrate the relationship between finitely exchangeable arrays and finitely exchangeable sequences. We then derive sharp bounds on the total variation distance between distributions of finitely and infinitely exchangeable arrays.

1. Introduction

The study of invariance properties of probability distributions has long been the focus of statisticians and probabilists (De Moivre, 1756; de Laplace, 1820; De Finetti, 1931; Hewitt and Savage, 1955; Aldous, 1985; Austin and Panchenko, 2013). The most common type of invariance is the notion of independent and identically distributed distributions. It is fundamental for classical results such as the weak and strong laws of large numbers and central limit theorems. Relaxation of the statistical notion of independence in this definition to simply requiring that the order of a sequence of random variables does not affect their joint distribution gives rise to the notion of exchangeable random variable sequences. These variables are marginally identically distributed but they can exhibit different forms of dependence. Infinite exchangeability of a sequence of random variables is in turn related back to independent and identically distributed random variable sequences via a famous theorem due to De Finetti (1931, 1972). Considering an infinitely exchangeable sequence $X_1, X_2, \ldots$ of binary random variables the theorem states that there is a unique probability measure μ such that for all $n \geq 1$

$$\Pr(X_1 = e_1, \ldots, X_n = e_n) = \int p^{\sum e_i}(1-p)^{n-\sum e_i} d\mu(p). \tag{1}$$

Generalizations of this famous theorem to higher dimensional arrays were given independently by Hoover (1989) and Aldous (1981): For example consider an infinitely row–column weakly exchangeable binary matrix X. The distribution of such a matrix is invariant under the permutation of the row and column indices of the matrix by the same permutation. Intuitively, such a matrix can represent relational data, where the shared index set of the rows and columns represents an actor, while the entries of the matrix represent the existence of a relationship among the actors. The Aldous–Hoover

* Corresponding author.
E-mail addresses: volfovsky@fas.harvard.edu (A. Volfovsky), airoldi@fas.harvard.edu (E.M. Airoldi).

http://dx.doi.org/10.1016/j.spl.2016.02.013
0167-7152/

representation for this matrix is

$$X_{ij} = \mathbf{1}[W(U_i, U_j) \geq \lambda_{ij}],$$

where U_i and λ_{ij} are independent uniform (0, 1) random variables and W is a measurable symmetric function from $[0, 1]^2 \to [0, 1]$. Intuitively, the U_i can be interpreted as actor attributes while the λ_{ij} as pairwise attributes.

These results have served as foundation for many Bayesian methods. For example, proofs of the consistency of Bayesian procedures rely on the assumption that the data is infinitely exchangeable, that is conditionally i.i.d. (Berk, 1970). Similarly, the analysis of relational data heavily relies on the Aldous–Hoover representation for arrays (Hoff et al., 2002). The power of these results is derived from the assumption that the index set of the array is infinite. Specifically, it is well known that de Finetti's theorem does not hold for finitely exchangeable sequences (Diaconis, 1977; Diaconis and Freedman, 1980). A similar result has been known, though apparently not written down, for the failure of the Aldous–Hoover representation to hold for finitely row–column exchangeable arrays. Consider the following distribution on 2×2 binary matrices: $P(x_{11} = x_{22} = 0) = 1$, $P(x_{12} = x_{21}) = 0$ and $P(x_{12} = 1) = P(x_{21} = 1) = 1/2$. This is clearly a jointly row–column exchangeable distribution but an Aldous–Hoover representation would require:

$$0 = P(x_{12} = 0, x_{21} = 0) = \int (1-p)^2 d\mu(p)$$

$$0 = P(x_{12} = 1, x_{21} = 1) = \int p^2 d\mu(p)$$

for μ a mixing measure as in the de Finetti representation in Eq. (1).

Notions of finite invariance described above have received a lot of attention in the case of sequences, but much less so for networks and general arrays. Specifically, our work is motivated by the pervasiveness of the assumption of joint row–column exchangeability throughout the field of network analysis (Holland et al., 1983; Hoff et al., 2002; Diaconis and Janson, 2007; Airoldi et al., 2009, 2013; Bickel and Chen, 2009). While certain network data have the potential representation as a sample from an infinite population, it is very common to observe the whole finite network such as in the study of trade between countries (Volfovsky and Hoff, 2015), friendship networks among students in the same class (Hoff et al., 2013) and interactions among monks in a monastery (Sampson, 1969). In these cases an assumption of exchangeability of the nodes is still desirable as there is no information in the individual labels, but the assumption of infinite exchangeability is inappropriate since there is no infinite network the data could have been sampled from.

This article is motivated by these finitely exchangeable network scenarios and is not intended as a review of infinite exchangeability. A comprehensive overview of the assumptions and representations of infinite exchangeability for networks and general arrays is available in Orbanz and Roy (2013). In the next section we derive sharp bounds on the total variation distance between finitely exchangeable network distributions and their infinitely exchangeable extensions. This is done by relating the distribution of the singular value decomposition of a relational matrix to the distribution of finitely exchangeable sequences. Section 3 extends these results to general k dimensional arrays where each dimension is independently exchangeable. To arrive at this result we extend the results of Freedman (1977) to bounds on the total variation distance between sampling with and without replacement from multiple urns. The desire for an exact bound was posed as open problem 15.10 by Aldous (1985, p.137). In the Discussion we provide an explicit statement of a representation theorem for distributions that are invariant under the operations of a finite group G. This representation can be used to compute a G-invariant non-parametric maximum likelihood estimate of a distribution as well as for the construction of test statistics for the infinite exchangeability of a distribution.

2. Sharp total variation bounds for networks

In this section we derive bounds for the total variation distance between the distributions of finitely and infinitely exchangeable networks. This quantitative summary provides insight into the efficacy of the assumption of infinite exchangeability. By taking limits of the bounds in the number of nodes in the network we recover the classical Aldous–Hoover representation. Recall that a distribution of a network with m nodes is exchangeable, that is its representation as a square relational data matrix X is jointly row–column exchangeable, if for all $g \in S_m$(=symmetric group) we have $\Pr(gX = A) = \Pr(X = A)$ where g acts on the rows and columns of X simultaneously. The results of Aldous (1981) and Hoover (1979) provide a representation theorem for infinitely exchangeable networks in the spirit of de Finetti's theorem for infinitely exchangeable sequences: If $(X_{ij})_{1,1}^{\infty,\infty}$ is an infinitely exchangeable network then it can be written as $X_{ij} = f(\alpha, u_i, u_j, \lambda_{ij})$ for a measurable function f that is symmetric in its middle arguments and $\alpha, u_i, \lambda_{ij}$ all independent uniform (0, 1). This is explicitly called joint or weak exchangeability. This representation is frequently used when describing relational data where the object of interest is a square array X.

It is natural to ask how close the finitely exchangeable distributions are to the Aldous–Hoover representation of an infinitely row–column exchangeable array. The example of the failure of the Aldous–Hoover representation for jointly exchangeable arrays is similar in spirit to the classical example provided by Diaconis (1977) for the failure of the de Finetti's theorem for a finite sequence. We first present the sharp total variation bound for finitely exchangeable sequences found in Diaconis and Freedman (1980) and then employ a singular value decomposition to extend the result to networks.